Bayesian probability theory and generative models

Bruno A. Olshausen[*]

November 8, 2006

Abstract

Bayesian probability theory provides a mathematical framework for peforming inference, or reasoning, using probability. The foundations of Bayesian probability theory were laid down some 200 years ago by people such as Bernoulli, Bayes, and Laplace, but it has been held suspect or controversial by modern statisticians. The last few decades though have seen the occurrence of a "Bayesian revolution," and Bayesian probability theory is now commonly employed (oftentimes with stunning success) in many scientific disciplines, from astrophysics to neuroscience. It is most often used to judge the relative validity of hypotheses in the face of noisy, sparse, or uncertain data, or to adjust the parameters of a specific model. Here we discuss the basics of Bayesian probability theory and show how it has been utilized in models of cortical processing and neural decoding.

Bayes' rule

Bayes' rule really involves nothing more than the manipulation of conditional probabilities. Remember that the joint probability of two events, $A\&B$, can be expressed as

$$\begin{aligned} P(AB) &= P(A|B)P(B) && (1) \\ &= P(B|A)P(A) && (2) \end{aligned}$$

In Bayesian probability theory, one of these "events" is the hypothesis, H, and the other is data, D, and we wish to judge the relative truth of the hypothesis given the data. According to Bayes' rule, we do this via the relation

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \qquad (3)$$

The term $P(D|H)$ is called the *likelihood* function and it assesses the probability of the observed data arising from the hypothesis. Usually this is known by the

[*] Much of this material is adapted from the excellent treatise by T.J. Loredo, "From Laplace to supernova SN 1987A: Bayesian inference in astrophysics" in *Maximum entropy and Bayesian methods*, Kluwer, 1989.

experimenter, as it expresses one's knowledge of how one expects the data to look given that the hypothesis is true. The term $P(H)$ is called the *prior*, as it reflects one's prior knowledge before the data are considered. The specification of the prior is often the most subjective aspect of Bayesian probability theory, and it is one of the reasons statisticians held Bayesian inference in contempt. But closer examination of traditional statistical methods reveals that they all have their hidden assumptions and tricks built into them. Indeed, one of the advantages of Bayesian probability theory is that one's assumptions are made up front, and any element of subjectivity in the reasoning process is directly exposed. The term $P(D)$ is obtained by integrating (or summing) $P(D|H)P(H)$ over all H, and usually plays the role of an ignorable normalizing constant. Finally, the term $P(H|D)$ is known as the *posterior*, and as its name suggests, reflects the probability of the hypothesis after consideration of the data.

Another way of looking at Bayes' rule is that it represents learning. That is, the transformation from the prior, $P(H)$, to the posterior, $P(H|D)$, formally reflects what we have learned about the validity of the hypothesis from consideration of the data. Now let's see how all of this is played out in a rather simplified example.

A simple example

A classic example of where Bayesian inference is employed is in the problem of estimation, where we must guess the value of an underlying parameter from an observation that is corrupted by noise. Let's say we have some quantity in the world, x, and our observation of this quantity, y, is corrupted by additive Gaussian noise, n, with zero mean:

$$y = x + n \tag{4}$$

Our job is to make the *best guess* as to the value of x given the observed value y. If we knew the probability distribution of x given y, $P(x|y)$, then we might want to pick the value of x that maximizes this distribution

$$\hat{x} = \arg\max_x P(x|y)\,. \tag{5}$$

Alternatively, if we want to minimize the mean squared error of our guesses, then we should pick the mean of $P(x|y)$:

$$\hat{x} = \int x\, P(x|y)\, dx\,. \tag{6}$$

So, if only we knew $P(x|y)$ then we could make an optimal guess.

Bayes' rule tells us how to calculate $P(x|y)$:

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}\,. \tag{7}$$

The two main things we need to specify here are $P(y|x)$ and $P(x)$. The first is easy, since we specified the noise, n, to be Gaussian, zero-mean, and additive. Thus,

$$P(y|x) \quad = \quad P(n + x|x) \tag{8}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(y-x)^2}{2\sigma_n^2}} \,, \tag{9}$$

where σ_n^2 is the variance of the noise. For the prior, we have to draw upon our existing knowledge of x. Let's say x is the voltage of a car battery, and as an experienced mechanic you have observed the voltages on thousands of cars, using very accurate voltage meters, to have a mean of 12 volts, variance of 1 volt, with an approximate Gaussian distribution. Thus,

$$P(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \,. \tag{10}$$

where $\mu_x = 12$ and $\sigma_x^2 = 1$. Now we are in a position to write down the posterior on x:

$$\begin{aligned} P(x|y) &\propto P(y|x)P(x) & (11) \\ &= e^{-\frac{(y-x)^2}{2\sigma_n^2}} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} & (12) \\ &= e^{-\frac{1}{2}\left[\frac{(y-x)^2}{\sigma_n^2} + \frac{(x-\mu_x)^2}{\sigma_x^2}\right]} \,. & (13) \end{aligned}$$

The x which maximizes $P(x|y)$ is the same as that which minimizes the exponent in brackets which may be found by simple algebraic manipulation to be

$$\hat{x} = \frac{\sigma_x^2\, y + \sigma_n^2\, \mu_x}{\sigma_x^2 + \sigma_n^2} \tag{14}$$

The entire inference process is depicted graphically in terms of the probability distributions in figure 1.

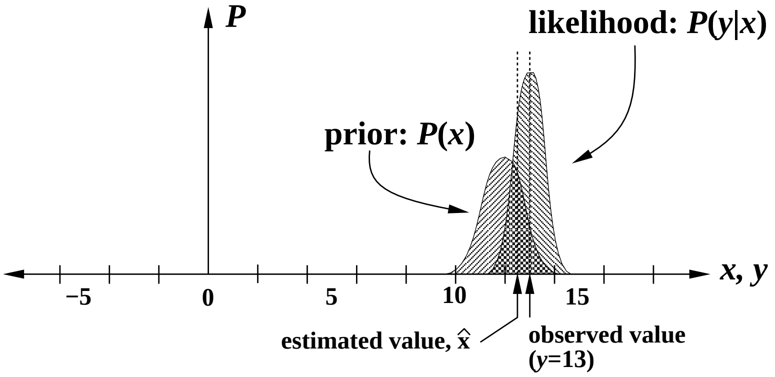

Figure 1: A simple example of Bayesian inference.

Generative models

Generative models, also known as "latent variable models" or "causal models," provide a way of modeling how a set of observed data could have arisen from a set of underlying causes. Such models have been commonly employed in the social sciences, usually in the guise of factor analysis, to make inferences about the causes leading to various social conditions or personality traits. More recently, generative models have been used to model the function of the cerebral cortex. The reason for this is that the cortex can be seen as solving a very complex inference problem, where it must select the best hypothesis for "what's out there" based on the massive data stream (gigabits per second) present on the sensory receptors.

The basic idea behind a generative model is illustrated in figure 2. A set of multivariate data, D, is explained in terms of a set of underlying causes, α. For instance, the data on the left may be symptoms of a patient, the causes on the right might be various diseases, and the links would represent how the diseases give rise to the symptoms and how the diseases interact with each other. Alternatively, the data may be a retinal image array, in which case the causes would be the objects present in the scene along with their positions and that of the lighting source, and the links would represent the rendering operation for creating an image from these causes. In general the links may be linear (as is the case in factor analysis), or more generally they may instantiate highly non-linear interactions among the causes or between the causes and the data.

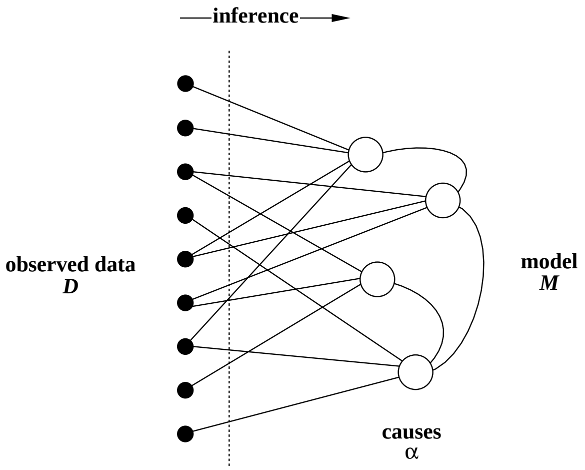

Figure 2: A generative model.

There are two fundamental problems to solve in a generative model. One is to infer the best set of causes to represent a *specific* data item, D_i. The other is to learn the best model, M, for explaining the *entire* set of data, $D = \{D_1, D_2, ..., D_n\}$. For example, each D_i might correspond to a specific image, and D would represent the set of all observed images in nature. In modeling the operations of the cortex, the first problem may be seen as one of *perception*, while the second is one of *adaptation*.

Inference (perception)

Inferring the best set of causes to explain a given piece of data usually involves maximizing the posterior over α (or alternatively computing its mean), similar to the computation of x in the simplified example above.

$$\begin{aligned} \hat{\alpha} &= \arg\max_{\alpha} P(\alpha|D_i, M) && (15) \\ &= \arg\max_{\alpha} P(D_i|\alpha, M) P(\alpha|M)\,. && (16) \end{aligned}$$

Note that the denominator $P(D_i|M)$ may be dropped here since D_i is fixed. All quantities are conditioned on the model, M, which specifies the overall "architecture" (such as depicted in the figure) within which the causes, α, are defined.

Learning (adaptation)

The model, M, specifies the set of potential causes, their prior probabilities, and the generative process by which they give rise to the data. Learning a specific model, M, that best accounts for all the data is accomplished by maximizing the posterior distribution over the models, which according to Bayes' rule is

$$P(M|D) \propto P(D|M) P(M)\,. \tag{17}$$

Oftentimes though we are agnostic in the prior over the model, and so we may simply choose the model that maximizes the likelihood, $P(D|M)$. The total probability of all the data under the model is

$$\begin{aligned} P(D|M) &= P(D_1|M) \times P(D_2|M) \times ... \times P(D_n|M) && (18) \\ &= \Pi_i P(D_i|M) && (19) \end{aligned}$$

where D_i denotes an individual data item (e.g., a particular image). The probability of an individual data item is obtained by summating over all of the possible causes for the data

$$P(D_i|M) = \sum_{\alpha} P(D_i|\alpha, M) P(\alpha|M)\,. \tag{20}$$

It is this summation that forms the most computationally formidable aspect of learning in the Bayesian framework. Usually there are ways we can approximate this sum though, and much effort goes into this step.

One typically maximizes the log-likelihood of the model for reasons stemming from information theory as well as the fact that many probability distributions are naturally expressed as exponentials. In this case, we seek a model, M^*, such that

$$\begin{aligned} M^* &= \arg\max_{M} \log P(D|M) && (21) \\ &= \arg\max_{M} \sum_i \log P(D_i|M) && (22) \\ &= \arg\max_{M} \langle \log P(D|M) \rangle\,. && (23) \end{aligned}$$

An example of where this approach has been applied in modeling the cortex is in understanding the receptive field properties of so-called simple cells, which are found in the primary visual cortex of all mammals. These cells have long been noted for their spatially localized, oriented, and bandpass receptive fields, but until recently there has been no explanation, in terms of a single quantitative theory, for why cells are built this way. In terms of Bayesian probability theory, one can understand the function of these cells as forming a model of natural images based on a linear superposition of sparse, statistically independent events. That is, if one utilizes a linear generative model where the prior on the causes is factorial and sparse, the set of linear weighting functions that emerge from the adaptation process (23) are localized, oriented, and bandpass, similar to the functions observed in cortical cells and also to the basis functions of commonly used wavelet transforms (Olshausen & Field, 1996, *Nature*, 381:607-9).

Neural decoding

Another problem in neuroscience where Bayesian probability theory has proven useful is in the decoding of spike trains. One of the most common experimental paradigms employed in sensory neurophysiology is to record the activity of a neuron while a stimulus is being played to the animal. Usually the goal in these experiments is to understand the function served by the neuron in mediated perception. One plays a stimulus, s, and observes a spike train, t, and the goal is to infer what t says about s. In terms of Bayesian probability theory, we can state the problem as

$$P(s|t) \propto P(t|s)P(s)\,. \tag{24}$$

This is similar to the problem faced by the nervous system—i.e., given messages received in the form of spike trains, it must make inferences about the actual signals present in the world. Bialek and colleagues pioneered the technique of stimulus reconstruction, which allows one to assess via this relation (24) the amount of information relayed by a neuron about a particular stimulus. Equation 24 demonstrates an important use of the prior in Bayesian probability theory, as it makes explicit the fact that optimal inferences about the world, given the information in the spike train, depend on what assumptions are made about the structure of the world, $P(s)$. More likely than not, $P(s)$ is crucial for neurons in making inferences about the world, and so it is also important for experimentalists to consider the influence of this term when trying to ascertain the function of neurons from their responses to controlled stimuli.